# Elements of a New Ethical Framework for Big Data Research

**Washington and Lee Law Review Online**

Volume 72 | Issue 3 Article 5

3-31-2016

# Elements of a New Ethical Framework for Big Data Research

Effy Vayena
*University of Zurich*

Urs Gasser
*Harvard Law School*

Alexandra Wood
*Harvard University*

David R. O'Brien
*Harvard University*

Micah Altman
*Massachusetts Institute of Technology*

# Elements of a New Ethical Framework for Big Data Research

Effy Vayena, Urs Gasser, Alexandra Wood, David R. O'Brien, and Micah Altman*

*Emerging large-scale data sources hold tremendous potential for new scientific research into human biology, behaviors, and relationships. At the same time, big data research presents privacy and ethical challenges that the current regulatory framework is ill-suited to address. In light of the immense value of large-scale research data, the central question moving forward is not whether such data should be made available for research, but rather how the benefits can be captured in a way that respects fundamental principles of ethics and privacy.*

---

* Effy Vayena, Swiss National Science Foundation Professor & Division Head, Health Ethics and Policy Lab, Institute of Epidemiology, Biostatistics and Prevention, University of Zurich.

Urs Gasser, Professor of Practice, Harvard Law School; Executive Director, Berkman Center for Internet & Society, Harvard University.

Alexandra Wood, Fellow, Berkman Center for Internet & Society, Harvard University.

David R. O'Brien, Senior Researcher, Berkman Center for Internet & Society, Harvard University.

Micah Altman, Director of Research, MIT Libraries, Massachusetts Institute of Technology; Non-Resident Senior Fellow, The Brookings Institution.

The authors describe contributions to this Essay using a standard taxonomy. *See* Liz Allen, Jo Scott, Amy Brand, Marjorie Hlava & Micah Altman, *Publishing: Credit Where Credit Is Due*, 508 NATURE 312, 312–13 (2014). Vayena provided the core formulation of the Essay's goals and aims, and Wood led the writing of the original manuscript. All authors contributed to conceptualization through additional ideas and through commentary, review, editing, and revision. This material is based upon work supported by the National Science Foundation under Grant No. 1237235. Vayena is supported by the Swiss National Science Foundation. The manuscript was prepared for the *Beyond IRBs: Designing Ethical Review Processes for Big Data Research* workshop, held by the Future of Privacy Forum in Washington, D.C., on December 10, 2015. The authors wish to thank the Future of Privacy Forum and the workshop attendees for their contributions at the workshop, as well as their colleagues through the Privacy Tools for Sharing Research project at Harvard University for articulating ideas that underlie many of the conclusions drawn in this Essay.

*In response, this Essay outlines elements of a new ethical framework for big data research. It argues that oversight should aim to provide universal coverage of human subjects research, regardless of funding source, across all stages of the information lifecycle. New definitions and standards should be developed based on a modern understanding of privacy science and the expectations of research subjects. In addition, researchers and review boards should be encouraged to incorporate systematic risk-benefit assessments and new procedural and technological solutions from the wide range of interventions that are available. Finally, oversight mechanisms and the safeguards implemented should be tailored to the intended uses, benefits, threats, harms, and vulnerabilities associated with a specific research activity.*

*Development of a new ethical framework with these elements should be the product of a dynamic multistakeholder process that is designed to capture the latest scientific understanding of privacy, analytical methods, available safeguards, community and social norms, and best practices for research ethics as they evolve over time. Such a framework would support big data utilization and help harness the value of big data in a sustainable and trust-building manner.*

*Table of Contents*

I. Introduction ..... 422

II. Recent Illustrations of Oversight Issues in Big Data Research ..... 424

III. Gaps in the Scope of the Existing Regulatory Framework ..... 425

IV. The Inadequacy of Informed Consent Requirements ..... 431

V. Recommendations for a New Ethical Framework for Big Data Research ..... 432

A. Universal Coverage ..... 434
B. Conceptual Clarity ..... 435
C. Risk-Benefit Assessments ..... 436
D. New Procedural and Technological Solutions ..... 437
E. Tailored Oversight ..... 438

VI. Multistakeholder Process for the Development of a Framework ..... 439

VII. Conclusions ..... 441

## *I. Introduction*[1]

Vast quantities of data about individuals are increasingly being created by new services such as mobile apps and through methods such as DNA sequencing.[2] These data sources can be quite rich, containing large numbers of fine-grained data points related to human biology, behaviors, and relationships over time.[3] Because they can enable analyses at an unprecedented level of detail, these large-scale data sources hold tremendous potential for scientific inquiry. In addition, the costs of obtaining, storing, and analyzing data from these sources are low and continuing to

1. This Essay summarizes, in part, joint work with other collaborators. Jeffrey P. Kahn, Effy Vayena & Anna C. Mastroianni, *Learning As We Go: Lessons from the Publication of Facebook's Social-Computing Research*, 111 PROCEEDINGS OF THE NAT'L ACAD. OF SCIS. 13677 (2014); Micah Altman, Alexandra Wood, David R. O'Brien, Salil Vadhan & Urs Gasser, *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L.J. (forthcoming 2016); Salil Vadhan et al., Comments to the Department of Health and Human Services Re: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators, Docket No. HHS-OPHS-2011-0005 (Oct. 26, 2011), *archived at* https://perma.cc/CK7V-V4AT; Effy Vayena, Marcel Salathé, Lawrence C. Madoff & John S. Brownstein, *Ethical Challenges of Big Data in Public Health*, 11 PLOS COMPUTATIONAL BIOLOGY e1003904 (2015); David R. O'Brien et al., *Integrating Approaches to Privacy Across the Research Lifecycle: When Is Information Purely Public?*, Berkman Ctr. Res. Pub. No. 2015-7 (2015), https://dash.harvard.edu/handle/1/16140637; Alexandra Wood et al., Comments to the Department of Health and Human Services Re: Federal Policy for the Protection of Human Subjects; Proposed Rules, Docket No. HHS–OPHS–2015–0008 (Jan. 6, 2016), *archived at* https://perma.cc/6JHM-X7YJ.

2. *See* PRESIDENT'S COUNCIL OF ADVISORS ON SCI. AND TECH., EXEC. OFFICE OF THE PRESIDENT, BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE, at ix (2014), https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf (describing how the "ubiquity of computing and electronic communication technologies has led to the exponential growth of data from both digital and analog sources").

3. *See id.* § 2.1, at 11–13 (providing examples of the types of data collected from new big data sources).

fall, relative to the costs of conducting traditional research studies.

For these reasons, big data are driving rapid advances in research, particularly through the emergence of fields such as computational social science and biomedical big data research.[4] Public health researchers, for example, are currently exploring ways to supplement traditional methods of disease outbreak detection by analyzing streams of data from social networks, chat rooms, and web search queries.[5] Looking ahead, interest in the research potential of big data is expected to continue to rise as the number of large-scale data sources increases and the technological capabilities for big data analysis improve.

We recognize the immense research value of big data and believe new large-scale data sources should be made available so that their full potential can be realized. At the same time, big data research presents new risks that the current regulatory framework is ill-suited to address. In light of the substantial value of large-scale data, the central question moving forward is not whether such data should be made available for research, but rather how the benefits can be captured in a way that respects fundamental principles of ethics and privacy. This Essay therefore recommends updates to the oversight framework that would help enable the collection, use, and sharing of big data in

---

4. *See* David Lazer et al., *Life in the Network: The Coming Age of Computational Social Science*, 323 SCIENCE 721, 721 (2009) (describing the adoption of computational social science methods by Internet companies, such as Google and Yahoo, and government agencies, like the U.S. National Security Agency); Gary King, *Ensuring the Data Rich Future of the Social Sciences*, 331 SCIENCE 719, 719–20 (2011) (providing an overview of how "[m]assive increases in the availability of informative social science data are making dramatic progress possible in analyzing, understanding, and addressing many major societal problems"); Eric Bender, *Big Data in Biomedicine*, 527 NATURE S1, S1 (2015) (providing a short update on developments in the use of big data in the field of biomedicine).

5. *See* Vayena et al., *supra* note 1, at 3 (discussing how big data sources are utilized in digital disease detection); *see also, e.g.*, Amy Wesolowski et al., *Quantifying the Impact of Human Mobility on Malaria*, 338 SCIENCE 267, 268 (2012) (describing the use of mobile phone data to track human travel and estimate its contribution to the spread of malaria in Kenya); David Talbot, *African Bus Routes Redrawn Using Cell-Phone Data*, MIT TECH. REV. (Apr. 30, 2013), https://www.technologyreview.com/s/514211/african-bus-routes-redrawn-using-cell-phone-data/ (reporting on the use of mobile phone data to optimize an urban transportation system in Ivory Coast).

line with ethical principles, research community norms, and the expectations of human subjects. Achieving this balance will be critical to ensuring the trust and support of the public and, ultimately, the long-term viability of big data research.

## *II. Recent Illustrations of Oversight Issues in Big Data Research*

There have recently been a number of high-profile incidents illustrating gaps in the oversight of big data research. Most notably, researchers involved in a joint Facebook-Cornell University study generated controversy in 2014 when they published the results of empirical research involving interventions with Facebook users without their knowledge.[6] The study aimed to observe changes in behavior and mood in response to variations in emotionally charged content viewed by users of the Facebook social media platform. These types of interventions almost certainly would have required approval from an institutional review board (IRB) had the research been conducted under a federal grant, rather than in a commercial setting.[7] This is just one example of the types of research activities increasingly conducted beyond the reach of traditional oversight due to the limited scope of the regulations in place.[8]

Potential oversight gaps have been discovered not only in study design and data collection but also in data release. In 2008, researchers published findings on a methodology for determining whether data about a specific individual are contained in a database including mixtures of genomic DNA collected from hundreds of people.[9] Although some believed the genomic DNA

---

6. Adam D. I. Kramer, Jamie E. Guillory & Jeffrey T. Hancock, *Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks*, 111 PROCEEDINGS OF THE NAT'L ACAD. OF SCIS. 8788 (2014).

7. For a detailed discussion of the applicability of the Common Rule to the emotional contagion experiment and social media studies more generally, see James Grimmelmann, *The Law and Ethics of Experiments on Social Media Users*, 13 COLO. TECH. L.J. 219 (2015).

8. *See generally id.* (discussing the current regulatory framework and recommending changes in light of big data research activities).

9. *See* Nils Homer et al., *Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays*, 4 PLOS GENETICS e1000167 (2008).

databases to be sufficiently aggregated so as to pose little risk to individual privacy, a group of researchers showed that an individual's participation in a study about a specific medical condition could potentially be confirmed using the released data. The National Institutes of Health revoked public access to two DNA databases as a result of this study, and other organizations that maintain similar databases are following suit.[10] In another study, researchers even demonstrated the potential to infer the surnames of individuals in de-identified genomic databases.[11]

More generally, privacy is a significant challenge for large-scale datasets, as the number of data points associated with a given record makes it highly likely for it to be unique and, therefore, identifiable.[12] Techniques for learning about individuals in a data release are rapidly advancing, enabling new scientific discoveries but also exposing vulnerabilities in many commonly used measures for protecting privacy. These vulnerabilities are calling into question regulatory approaches that broadly permit the public release of aggregated or de-identified data without the use of additional controls.

### *III. Gaps in the Scope of the Existing Regulatory Framework*

Human subjects research protection frameworks developed in the late 1970s fail to address many of the oversight challenges in big data research. Broadly speaking, social, behavioral, and

---

10. *See* Jason Felch, *DNA Profiles Blocked from Public Access*, L.A. TIMES, Aug. 29, 2008, http://articles.latimes.com/2008/aug/29/local/me-dna29 (reporting on revocations of public access to DNA databases due to privacy concerns).

11. *See* Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 SCIENCE 321 (2013).

12. *See* Yves-Alexandre de Montjoye et al., *Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata*, 347 SCIENCE 536, 536 (2015) (demonstrating that knowing the dates and locations of four purchases is sufficient to identify 90% of the people in a dataset of credit card transactions that has been stripped of information typically considered to be personally identifying); *see generally* Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, 3 NATURE SCI. REPS. (2013), http://www.nature.com/articles/srep01376 (finding that "in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals").

educational researchers have argued that the current regulatory framework emphasizes practices, such as obtaining informed consent and balancing the benefits of research against the risks of participation, that are out of place in non-clinical research.[13] These gaps are especially pronounced with respect to many types of big data research.[14] For example, when using data originally collected by a third party such as Facebook, a researcher has not interacted with the subjects of the data and informed them of the risks associated with their participation. Furthermore, regulations currently emphasize risk mitigation at the study design and data collection stages of the information lifecycle and, to a much lesser extent, those that arise in later stages, such as the transformation, dissemination, and post-access stages. Consequently, as advances in big data drive increased data sharing and re-use by researchers, more of their activities will be subject to limited or, in some cases, no oversight.

Definitions found in the federal policy for the protection of human subjects, known as the Common Rule,[15] also create gaps in oversight. What qualifies as human subjects research—and therefore falls within the purview of the Common Rule and IRB review—is rather narrowly defined.[16] Its scope is limited to research involving "a living individual about whom an investigator (whether professional or student) conducting research obtains (1) [d]ata through intervention or interaction with the individual, or (2) [i]dentifiable private information."[17] Many types of research conducted today using big data do not fall

---

13. *See* NAT'L RESEARCH COUNCIL, PROPOSED REVISIONS TO THE COMMON RULE: PERSPECTIVES OF SOCIAL AND BEHAVIORAL SCIENTISTS: WORKSHOP SUMMARY 10–12 (2013) (discussing critiques of the requirements of the Common Rule as applied in social, behavioral, and educational research).

14. *See* Letter from the Secretary's Advisory Committee on Human Research Protections to the Health and Human Services Secretary, Sylvia M. Burwell, Attachment A: Human Subjects Research Implications of "Big Data" Studies (Apr. 24, 2015).

15. Federal Policy for the Protection of Human Subjects, 7 C.F.R. pt. 1c, 10 C.F.R. pt. 745, 14 C.F.R. pt. 1230, 15 C.F.R. pt. 27, 16 C.F.R. pt. 1028, 24 C.F.R. pt. 60, 28 C.F.R. pt. 46, 32 C.F.R. pt. 219, 34 C.F.R. pt. 97, 38 C.F.R. pt. 16, 40 C.F.R. pt. 26, 45 C.F.R. pt. 46, 45 C.F.R. pt. 690, 49 C.F.R. pt. 11 (2015).

16. For a discussion of the gaps created by the Common Rule's definition of human subjects research, see generally Jeffrey P. Kahn et al., *supra* note 1.

17. 45 C.F.R. § 46.102(f) (2015).

squarely within this definition. For example, research using a pre-existing Facebook dataset arguably falls outside the scope of this definition because it does not involve an intervention or interaction between the researcher and the research subjects.[18]

The second part of this definition likely excludes from oversight some research associated with non-minimal risk of harm. For instance, it permits a researcher who conducts secondary analysis using a de-identified dataset to apply for an exemption from IRB review.[19] De-identification alone, however, does not minimize all privacy risks to subjects or necessarily protect personal information in the manner that most individuals would expect. A research dataset that has been de-identified can, in many cases, be re-identified easily.[20] For example, numerous attacks on de-identified datasets have demonstrated that it is often possible to identify individuals in data that have been stripped of direct and indirect identifiers.[21] It has been shown more generally that very few pieces of information can be used to uniquely identify an individual in a released set of data.[22]

As illustrated by the genomic DNA database examples provided above, data stripped of identifiers or released in aggregate form may nevertheless carry privacy risks. Alternatives to traditional de-identification techniques, such as privacy-aware methods for producing contingency tables, synthetic data, data visualizations, interactive mechanisms, and multiparty computations can in many cases provide strong guarantees of privacy while also largely preserving the utility of

---

18. For a discussion of this definition in the context of social media research, see Grimmelmann, *supra* note 7.

19. *See* 45 C.F.R. § 46.101(b)(4) (2015) (exempting "[r]esearch involving the collection or study of existing data . . . if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects").

20. *See* Arvind Narayanan & Edward W. Felten, *No Silver Bullet: De-Identification Still Doesn't Work*, at 1 (July 9, 2014), http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf, *archived at* https://perma.cc/RD6U-W4X7 (arguing "there is no evidence that de-identification works either in theory or in practice").

21. *See id.* (discussing numerous successful demonstrations of the potential to identify individuals in datasets that had been deemed de-identified).

22. *See* sources cited *supra* note 12 (demonstrating that as few as two to four data points can be sufficient to uniquely identify individuals in large-scale datasets).

the data.[23] Rather than promoting the use of more robust approaches such as these, however, the Common Rule arguably encourages the wide use and sharing of data that have been de-identified using heuristic techniques and released in forms that may be vulnerable to significant privacy risks.

The second part of the Common Rule's definition of human subjects research also exempts research using information considered to be non-private. The distinction between public and private information, however, is the subject of significant debate.[24] Sensitive information is increasingly captured in big data scraped from the web or observed via sensors in public spaces and used for research, often with little or no oversight.[25] Although the protections of the Common Rule apply to research using personal information that subjects have a reasonable expectation will not be made public, many individuals have mismatched expectations regarding secondary uses of information deemed to be public.[26] Consequently, some

23. *See* Salil Vadhan et al., *supra* note 1. Many of these advanced methods are also compatible with a strong, formal guarantee of privacy known as differential privacy. *See generally* Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 COMMC'NS ACM 86 (2011) (defining the framework of differential privacy).

24. For a discussion of the evolving notion of public information, see generally David R. O'Brien et al., *supra* note 1.

25. *See* ALEX PENTLAND, SOCIAL PHYSICS: HOW GOOD IDEAS SPREAD—THE LESSONS FROM A NEW SCIENCE 8–10 (2014) (exploring how research using big data collected from smartphones, GPS devices, and online platforms can yield insights into social behavior); R. Benjamin Shapiro & Pilar N. Ossorio, *Regulation of Online Social Network Studies*, 339 SCIENCE 144, 144 (2013) (describing research conducted on social networking web sites and the lack of guidance on conducting such research ethically); Michael Zimmer, *"But the Data Is Already Public": On the Ethics of Research in Facebook*, 12 ETHICS INFO. TECH. 313, 314 (2010) (discussing ethical concerns related to research using data from social networking web sites).

26. *See* Mary Madden, *Privacy Management on Social Media Sites*, PEW RESEARCH CTR. (Feb. 24, 2012), http://www.pewinternet.org/2012/02/24/privacy-management-on-social-media-sites, *archived at* https://perma.cc/QZ29-4GAU (finding that individuals may share personal information through social media as a result of a lack of understanding regarding how such information is retained and used by such services); Yabing Liu et al., *Analyzing Facebook Privacy Settings: User Expectations vs. Reality*, PROCEEDINGS OF THE 2011 ACM SIGCOMM CONFERENCE ON INTERNET MEASUREMENT 61, 63–65 (2011), http://conferences.sigcomm.org/imc/2011/docs/ p61.pdf (describing the results of an experiment highlighting the mismatch between Facebook's privacy practices and users' expectations).

commentators argue that IRBs and investigators should take steps to protect some personal information obtained from public sources.[27] Compare, for instance, the approach taken by the Common Rule to that found in the European Union, where many categories of information are protected as personal data despite their public nature.[28]

In response to this debate, research communities have developed ethical guidelines, and the Secretary's Advisory Committee on Human Research Protections has developed draft guidance on the use of data collected from Internet sources.[29] These resources aim to address many of the challenges associated with determining whether information collected online qualifies as public or private under the existing regulations.[30] However, what is considered to fall within these definitions is open to interpretation and will likely evolve over time. Further guidance on interpreting such standards and incorporating them into review board policies is needed.[31]

Another sizable subset of big data research not subject to the Common Rule is research supported exclusively by private funding. The sharp distinction between publicly and privately funded research results in inconsistent oversight, as many

---

27. For a discussion of proposals to address the gap between subject expectations and the use of publicly available data, see David R. O'Brien et al., *supra* note 1.

28. *See* Council Directive 95/46/EC, art. 2, 1995 O.J. (L. 281) 31 ("'[P]ersonal data' shall mean any information relating to an identified or identifiable natural person ('data subject')."); Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, at 41, COM (2012) 11 final (Jan. 25, 2012) ("'[P]ersonal data' means any information relating to a data subject.").

29. ANNETTE MARKHAM & ELIZABETH BUCHANAN, ETHICAL DECISION-MAKING AND INTERNET RESEARCH, RECOMMENDATIONS FROM THE AOIR ETHICS WORKING COMMITTEE (2012), http://aoir.org/reports/ethics2.pdf, *archived at* https://perma.cc/V2TR-4UA8; Letter from Secretary's Advisory Committee on Human Research Protections to the HHS Secretary, Attachment B: Considerations and Recommendations Concerning Internet Research and Human Subjects Research Regulations, with Revisions (May 20, 2013), http://www.hhs.gov/ohrp/sachrp/commsec/attachmentbsecletter20.pdf, *archived at* https://perma.cc/AB2P-UEC8.

30. *See* sources cited *supra* note 29.

31. *See* David R. O'Brien et al., *supra* note 1 (setting forth this proposition in greater detail).

privately funded research activities carry the same types of risks as research funded by the government.[32] In fact, identical studies conducted by two different organizations, one privately funded and another publicly funded, can be subject to markedly different requirements. Note, however, that institutional policies are evolving partially to address this gap. For example, many IRB policies cover certain research projects that are not federally funded,[33] and journal policies in many cases require all authors to undergo a formal ethical review before publication regardless of funding source.[34] A privately funded researcher may also come under the federal rules if she collaborates with a federally funded researcher. Furthermore, laws at the state level impose additional requirements for human subjects research protection that partially fill this gap.[35]

In addition to the various regulations and policies that apply to different classes of researchers within the United States, the regulations of foreign jurisdictions may also apply if any of the collaborating researchers or research subjects are located outside the United States. Many big data research initiatives are international in nature, and protections vary substantially depending on the national data protection regulation that applies. This can lead to mismatches between the safeguards

---

32. For a discussion of the gap in oversight for privately funded research, see Jeffrey P. Kahn et al., *supra* note 1.

33. *See, e.g.,* Policy: Commensurate Protections for Non-Federally Funded Human Subjects Research, UCLA OFFICE OF HUMAN RESEARCH PROT. PROGRAM, at 2 (June 4, 2013), http://ora.research.ucla.edu/OHRPP/Documents/Policy/10/CommensurateProtections.pdf, *archived at* https://perma.cc/24PP-RD76 ("The UCLA IRB applies protections equivalent to the Common Rule and Subparts A, B, C, and D to all non-federally funded research, with the following exceptions . . . .").

34. *See, e.g.*, Editorial Policies, BIOMED CENTRAL, https://www.biomedcentral.com/submissions/editorial-policies (last visited Feb. 28, 2016), *archived at* https://perma.cc/XV7V-QX3S ("Research involving human subjects, human material, or human data, must have been performed in accordance with the Declaration of Helsinki and must have been approved by an appropriate ethics committee.").

35. *See, e.g.,* N.Y. PUB. HEALTH LAW § 2442 (2016) ("No human research may be conducted in this state in the absence of the voluntary informed consent subscribed to in writing by the human subject."); MD. CODE ANN., HEALTH–GEN. § 13-2002(a) (West 2016) ("A person may not conduct research using a human subject unless the person conducts the research in accordance with the federal regulations on the protection of human subjects.").

used and the expectations and understanding of individual participants. For instance, research subjects may believe that the regulations of their home country protect their personal data, when in fact the requirements of another jurisdiction could be followed once their data cross a border.[36] The variations in treatment that result from the application of different regulatory requirements and expectations of privacy across jurisdictions creates challenges for researchers, particularly in the secondary analysis of data, as the location of every research subject might not be known. Furthermore, the fact that different protections may apply as research data about a subject moves between jurisdictions is generally not disclosed in consent forms. These factors contribute to uncertainty among researchers and subjects regarding which standards apply to a specific research activity, as well as overall inconsistency in research oversight.[37]

### *IV. The Inadequacy of Informed Consent Requirements*

Informed consent is a cornerstone of human subjects research protection. An approach based solely on notice and consent, however, has many known weaknesses. Consent forms and terms of service are lengthy, complex, and difficult to understand.[38] Disclosures often do not inform subjects of all potential data uses and the harms that could result from misuse of their personal information.[39] In addition, subjects generally have limited opportunities to withhold, revoke, or modify consent.

---

36. For a discussion of many of the issues that may arise when collecting, using, and sharing research data about human subjects across multiple jurisdictions, see David R. O'Brien et al., *supra* note 1.

37. *See* Jeffrey P. Kahn, *supra* note 1 (discussing inconsistencies in current oversight).

38. *See, e.g.*, S. Michael Sharp, *Consent Documents for Oncology Trials: Does Anybody Read These Things?*, 27 AM. J. CLINICAL ONCOLOGY 570 (2004); Carlos Jensen & Colin Potts, *Privacy Policies As Decision-Making Tools: An Evaluation of Online Privacy Notices*, *in* PROCEEDINGS OF THE SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS 471 (2004); Aleecia M. McDonald & Lorrie Faith Cranor, *The Cost of Reading Privacy Policies*, 4 I/S J. L. & POL'Y INFO. SOC'Y 560, 561 (2008).

39. *See* Irene Pollach, *What's Wrong with Online Privacy Policies?*, 50 COMMC'NS ACM 103, 103–08 (2007) (arguing that privacy policies tend not to build trust but rather exacerbate privacy concerns).

These concerns are heightened in big data research, which is often characterized by a substantial separation between the researcher and the research subject. For example, mobile and social networking platforms often embed notice about data collection and sharing practices, including the potential research uses of data collected through the platform, in terms of service, and individuals impliedly consent to sharing their data under such terms through their use of the service. Because the details are often buried in lengthy terms of service, users are likely unaware that they are participating in human subjects research through their use of a mobile or social networking platform alone.[40] More generally, the reliance on terms of service that are often vague, complex, and subject to modification without notice leaves users with an incomplete understanding of how their personal information will be used and shared by the service. These practices arguably fall short of the informed consent requirements intended by research ethics and regulatory frameworks that were developed for clinical research and the extensive recruitment and informed consent processes established in that context. If the research oversight framework is to be expanded to provide coverage for new categories of big data research, protections beyond the consent practices currently in wide use will likely be necessary.

## *V. Recommendations for a New Ethical Framework for Big Data Research*

A robust oversight framework is essential to safeguarding the interests of research subjects; ensuring trust, transparency, and accountability in the research community; maintaining

40. *See* Effy Vayena, Ann Mastroianni & Jeffrey Kahn, *Caught in the Web: Informed Consent for Online Health Research*, 5 SCI. TRANSLATIONAL MED. 173fs6, at 2 (2013), http://stm.sciencemag.org/content/scitransmed/5/173/173fs6.full.pdf ("[N]o publicly available studies have yet documented whether users understand or are even aware of the potential uses of their data when they access a site."); Jeffrey P. Kahn et al., *supra* note 1, at 13678 ("Unlike in psychology research, however, participants in social-computing studies may not be recruited in the usual sense, and so may not even realize they are participating in research, let alone that there may be interventions, including manipulation or deception, involved.").

continued support for, funding of, and participation in research studies; and realizing the full research potential of big data. As demonstrated by the gaps in oversight discussed above, changes to the existing framework are needed to continue to advance these values in big data research. At the core of this Essay's recommendations is recognition of both the ethical obligation to protect personal data[41] and the human right to participation in the production of scientific knowledge.[42] A component of the human right to science, the latter refers to the obligations of governments and other actors, including corporations,[43] to protect and promote participation in science across all stages of the research lifecycle.[44] An intervention designed to protect human subjects, therefore, should not prevent people who are willing to participate in a study from doing so and thereby impede the capacity of big data research to yield insights into human biology and behavior.

Below, this Essay provides a set of objectives and substantive components to consider as part of a new ethical framework guided by these values. In describing each objective of the proposed framework, this discussion also sketches example ways in which they could be met, through changes to regulations, the policies of review boards, and guidance on research community norms and industry best practices.

---

41. *See* European Data Protection Supervisor, *Towards a New Digital Ethics: Data, Dignity and Technology*, Opinion 4/2015 (Sept. 11, 2015) ("In today's digital environment, adherence to the law is not enough; we have to consider the ethical dimension of data processing.").

42. For a discussion of the human right to science and its application to the regulation of citizen science, see Effy Vayena & John Tasioulas, *"We the Scientists:" A Human Right to Citizen Science*, 28 PHIL. & TECH. 479 (2015).

43. Corporations are increasingly being called on to protect and respect human rights, *see* UNITED NATIONS HUMAN RIGHTS COUNCIL, GUIDING PRINCIPLES ON BUSINESS AND HUMAN RIGHTS: IMPLEMENTING THE UNITED NATIONS "PROTECT, RESPECT AND REMEDY" FRAMEWORK 32–33 (2011), http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf, including rights related to scientific progress.

44. *See id.* at 20 ("Because human rights situations are dynamic, assessments of human rights impacts should be undertaken at regular intervals: prior to a new activity or relationship; prior to major decisions or changes in the operation . . .; in response to or anticipation of changes in the operating environment.")

### *A. Universal Coverage*

Oversight should aim to cover the full scope of human subjects research. Changes to the existing framework are needed to address gaps in coverage for research involving many categories of information deemed to be public or non-identifiable, research that is privately funded, and research activities across all stages of the lifecycle, including the storage, processing, analysis, release, and post-release stages.[45] Encouraging the use of a wider range of privacy and security controls and moving towards the model adopted by several European countries, in which regulations cover all research activities regardless of the institution or source of funding,[46] are potential ways to address this gap.

To reduce the burden on IRBs as a result of an expanded scope of coverage, some responsibilities could be shared with emerging review bodies, such as consumer review boards,[47] participant-led review boards,[48] and personal data cooperatives.[49] For research subject to IRB review, regulators should consider adopting new exemptions to full review that are based in part on the risk-benefit determination described below, as well as explore emerging technological solutions for automating review decisions. In addition, changes to the Common Rule could direct IRBs to implement a limited review process for all research at the proposal stage, followed by regular monitoring throughout the

---

45. For an example framework for systematically analyzing privacy risks and intended data uses and aligning them with appropriate interventions at each stage of the information lifecycle, see generally Micah Altman et al., *supra* note 1.

46. *See* Council Directive 2001/20/EC, 2001 O.J. (L. 121) 34, http://ec.europa.eu/health/files/eudralex/vol-1/dir_2001_20/dir_2001_20_en.pdf, *archived at* https://perma.cc/UW8F-FQUQ.

47. *See* Ryan Calo, *Consumer Subject Review Boards: A Thought Experiment*, 66 STAN. L. REV. ONLINE 97, 101–02 (2013) (exploring the potential benefits of consumer subject review boards).

48. *See* Effy Vayena & John Tasioulas, *Adapting Standards: Ethical Oversight of Participant-Led Health Research*, 10 PLOS MED. e1001402 (2013).

49. *See* Ernst Hafen, Donald Kossmann & Angela Brand, *Health Data Cooperatives—Citizen Empowerment*, 53 METHODS INFO. MED. 82, 84 (2014); *see also* Effy Vayena & Urs Gasser, *Between Openness and Privacy in Genomics*, 13 PLOS MED. e1001937 (2016).

research lifecycle to identify research activities for which additional review is needed.

### *B. Conceptual Clarity*

Revised definitions and standards for privacy protection, as well as guidance on interpreting these definitions and applying appropriate safeguards, would likely help IRBs and investigators provide adequate and consistent protection for human subjects. As discussed above, the Common Rule's definition of human subjects research, particularly its reliance on a sharp binary determination based on the presence of "identifiable private information," leads to inconsistency and uncertainty in practice.[50] To provide clarity, the regulations should establish definitions for terms such as privacy, confidentiality, security, and sensitivity, and the terminology should be used consistently.[51]

Changes to the Common Rule could include language directing investigators to implement a combination of both security and privacy controls, where security controls can be viewed as restricting access to information, and privacy controls as limiting the potential for harm once access to information is granted. The regulations could also be revised to incorporate definitions based on a modern understanding of privacy that is not based on a strict binary conception of identifiability or public availability. For instance, the notion of privacy risk should cover more broadly the potential for others to learn about individuals based on the inclusion of their information in a set of data, as well as establish a privacy goal against which a technique for privacy protection can be evaluated.[52]

Regulators and review boards should consider consulting with ethics and privacy experts or establishing a regularly-convening advisory committee to provide concrete

---

50. For a more extensive discussion of the weaknesses of the Common Rule's binary identifiability standard, see Alexandra Wood et al., *supra* note 1, at 16.

51. For example definitions for these terms, see Micah Altman et al., *supra* note 1, at *30–31; Alexandra Wood et al., *supra* note 1, at 5–6.

52. *See* Alexandra Wood et al., *supra* note 1, at 5 (setting forth this proposal).

recommendations, as they formulate clarifying definitions, practices, methodologies, and guidelines for implementing up-to-date privacy practices. In particular, this expert body could help develop detailed guidance for review boards to reference as they incorporate revised concepts into their processes and the materials provided to researchers and research subjects. Regulators should also consider establishing a clearinghouse of review board policies and decisions that would enable the administrators of such bodies to learn from one another and achieve greater consistency in the application of standards for human subjects protection.

### *C. Risk-Benefit Assessments*

Researchers and review boards should be encouraged to incorporate systematic risk-benefit assessments.[53] Such assessments should evaluate the benefits that would accrue to society as a result of a research activity, the intended uses of the data involved, the privacy threats and vulnerabilities associated with the research activity, and the potential harms to human subjects as a result of the inclusion of their information in the data.[54] Results from this assessment can be used to guide the selection of protections that are calibrated to the specific risks and uses associated with a given research activity.[55]

Regulators, in consultation with data privacy experts, should consider developing detailed guidance to help review boards and researchers systematically examine the privacy threats and vulnerabilities at each stage of the information lifecycle, drawing from concepts found in the technical literature on data privacy

---

53. For an introduction to the components of such a risk-benefit assessment model, see the framework for a modern privacy analysis proposed in Micah Altman et al., *supra* note 1, at *29–57, and the reference guide for conducting benefit-risk analyses provided in Jules Polonetsky, Omer Tene & Joseph Jerome, *Benefit-Risk Analysis for Big Data Projects*, FUTURE OF PRIVACY FORUM, at 7–8 (Sept. 2014), https://fpf.org/wp-content/uploads/FPF_DataBenefitAnalysis_FINAL.pdf.

54. *See* Micah Altman et al., *supra* note 1, at *30–49 (outlining how to characterize and assess such risks and benefits).

55. *See id.* at *51–57 (proposing a framework for designing data releases "by aligning use, threats, and vulnerabilities with controls").

and information security.[56] An expert body could, for example, be involved in the development of guidance on modeling the risks and uses associated with a research activity, and selecting privacy controls that are aligned with these factors. Review boards could, in turn, use this general guidance as a basis for developing more detailed materials specific to their institutional contexts. As the nature of the benefits and risks changes over time, assessments should also evolve, and therefore regulators should consider consulting regularly with an expert body to update the guidance materials that are produced.

### *D. New Procedural and Technological Solutions*

Researchers should be incentivized to select from the wide range of procedural, economic, legal, educational, and technical protections that are available, rather than to rely on a narrow subset of controls, such as consent and de-identification.[57] Adoption of techniques from the full scope of available controls could be encouraged through revisions to the Common Rule requiring researchers to consider implementing reasonable and appropriate procedural, economic, legal, educational, or technical safeguards at each stage of the information lifecycle. In addition, regulatory language referring to consent and de-identification could be amended to acknowledge that in many cases these measures should be used in conjunction with additional controls, including information security controls.

Regulators should also consider creating a safe harbor for researchers who use robust privacy-preserving techniques.[58] Regulators, in consultation with an expert body of privacy researchers, IRB administrators, and researchers, could be authorized to compile a list of approved techniques that provide a strong guarantee of privacy protection. Examples of some of the technological controls that should be considered for inclusion in

---

56. For a framework for analyzing informational harms throughout the information lifecycle, see *id.* at *45–51.

57. For an expansive catalog of the privacy and security controls that researchers should be encouraged to consider adopting, see *id.* at 34–45.

58. For a proposal outlining a Common Rule safe harbor for certain privacy-preserving techniques, see Salil Vadhan et al., *supra* note 1, at 7–8.

such a list include privacy-aware methods for contingency tables, synthetic data, data visualizations, interactive mechanisms, and multiparty computations.[59] Revisions to the regulations could also require regulators and experts to meet regularly to update the list of approved techniques to reflect technological advances. Revised guidance materials could also cover new approaches to established controls such as consent, including methods for standardizing privacy policies for ease of understanding[60] and processes for dynamic consent that enable individuals to grant, modify, and revoke fine-grained research permissions over time.[61]

### *E. Tailored Oversight*

No one-size-fits-all solution to privacy exists, and researchers should instead be encouraged to adopt procedures and safeguards that are calibrated to the intended uses of the information collected; the benefits of the research activity; and the threats, vulnerabilities, and harms associated with the activity. One way to tailor oversight is to subject different categories of research activities to oversight by different review boards, including IRBs, consumer review boards, participant-led review boards, or personal data cooperatives. For example, in cases where IRB review is not required by the Common Rule, seeking approval from an appropriate review board could be required by journal editors or institutional policies and recommended more generally as an industry or research community best practice.

Oversight can also be tailored at the data sharing stage through tiered access.[62] Tiered access enables a data provider to

---

59. *See id.* at 4 (listing and defining these examples of privacy-preserving techniques).

60. *See* Lorrie Faith Cranor, *Necessary But Not Sufficient: Standardized Mechanisms for Privacy Notice and Choice*, 10 J. TELECOMM. & HIGH TECH. L. 273, 287–88, 305–06 (2012) (examining efforts to create standardized privacy notices and setting forth the benefits of standardization).

61. *See* Jane Kaye et al., *Dynamic Consent: A Patient Interface for Twenty-First Century Research Networks*, 23 EUR. J. HUM. GENETICS 141, 143 (2014) (explaining how dynamic consent platforms can be tailored to consumers' privacy preferences).

62. For a discussion of tiered access mechanisms and examples illustrating how they can be designed, see generally Salil Vadhan et al., *supra* note 1; Alexandra Wood et al., *supra* note 1; Micah Altman et al., *supra* note 1.

match the data-sharing mechanism with the risks of sharing such data, including factors related to the structure of the data, the sensitivity of the information and potential harms of disclosure, the level of consent obtained from subjects, the credentials of the intended recipients of the data, and the types of analyses they intend to perform.[63] For example, sharing aggregate data using one of the privacy-aware methods described above, such as statistics in the form of contingency tables generated using methods providing a formal guarantee of privacy, could be deemed a suitable option for making data available to the public. An intermediate level could allow approved researchers with proper credentials to analyze the data through a protected server after agreeing to the terms of a data use agreement, providing the data subjects with additional legal protections from misuse. For full access to raw data, individuals, such as academic researchers, could apply for access to the data through a monitored data environment, such as a virtual data enclave, under the terms of a data use agreement.

Similar mechanisms for aligning safeguards with intended uses can be implemented at other stages in the research lifecycle. For example, data minimization and purpose specification principles, operationalized through computable policies, could be applied at the study design and data collection stages to ensure that only the minimum amount of information is collected from human subjects and that data uses are restricted to those authorized by the subjects. Regulators, in consultation with data privacy experts, could establish guidance on tailoring controls at each stage of the lifecycle and implementing a tiered access mechanism. Additionally, review boards could be empowered to supplement this guidance with detailed instructions specific to their institutional contexts.

## *VI. Multistakeholder Process for the Development of a Framework*

Development of a new ethical framework with these components should be the product of a multistakeholder process,

63. For an example of a systematic framework for matching privacy interventions to the threats, harms, and vulnerabilities in a specific data release case, see Micah Altman et al., *supra* note 1, at *29–57.

with involvement from researchers, institutional review board administrators, industry representatives, regulators, ethicists, journal editors, and research subjects.[64] In addition to established principles of human subjects research protection, this multistakeholder group should be guided by the human right to science,[65] which includes the right to participate in the production of scientific knowledge, and seek to harmonize the latter right with other interests, such as the right to privacy. One output this group could consider developing is a set of ethical norms based in part on existing best practices for research ethics. A panel of domain experts from fields such as computer science, information security, law, and ethics could be convened to develop recommendations regarding practices, methodologies, and tools that are appropriate in different contexts, which could in turn inform the multistakeholder group's assessment of existing best practices. The set of norms developed by the group might begin as general guidelines but evolve over time into more formal codes of practice.

Interfacing with existing ethics and IRB processes, as well as with emerging oversight processes, such as consumer review boards, participant-led review boards,[66] and personal data cooperatives, would likely be a key component of this process. Regulators and institutional review board administrators, as stakeholders in this process, could evaluate the extent to which the current regulatory system is compatible with big data research, or whether changes to the Common Rule would be

---

64. For a discussion of a proposal to convene a multistakeholder group to develop ethical guidelines for big data research, see generally Effy Vayena & Urs Gasser, *Strictly Biomedical? Sketching the Ethics of the Big Data Ecosystem in Biomedicine*, *in* ETHICS OF BIOMEDICAL BIG DATA (Brent Mittelstadt & Luciano Floridi eds., forthcoming 2016).

65. *See* Farida Shaheed (Special Rapporteur in the Field of Cultural Rights), *Report on the Copyright Policy & the Right to Science & Culture*, U.N. Doc. A/HRC/28/57, at 4–5 (Dec. 24, 2014); Farida Shaheed (Special Rapporteur in the Field of Cultural Rights), *Report on the Right to Enjoy the Benefits of Scientific Progress and Its Applications*, U.N. Doc. A/HRC/20/26, at 3 (May 14, 2012); U.N. High Commissioner for Human Rights, *Report on the Seminar on the Right to Enjoy the Benefits of Scientific Progress and Its Applications*, U.N. Doc. A/HRC/26/19, at 5 (April 1, 2014).

66. *See* Effy Vayena et al., *Research Led by Participants: A New Social Contract for a New Kind of Research*, J. MED. ETHICS, at 1 (Mar. 30, 2015), http://jme.bmj.com/content/early/2015/03/30/medethics-2015-102663.full.pdf.

required. The multistakeholder group could also assess whether institutional review boards are appropriate as the primary oversight body for big data research. Alternatively, it may find that technological solutions can help automate some decisions traditionally made by IRBs, or that oversight by consumer review boards, participant-led review boards, or personal data cooperatives are better suited to the oversight of big data research.

Researchers should also be involved in the formulation of the framework, in recognition of the human right to participation in science across the entire lifecycle of research. Researcher input would likely help ensure that the oversight framework does not create new inefficiencies or burdens on the research process.

Finally, the multistakeholder group would likely benefit from regular meetings to review and update the framework once it is in place, to ensure its flexibility and adaptability to unforeseen technological advancements, emerging study design and analytical techniques, new research questions, evolving privacy and other risks to human subjects, regulatory shifts, and changes in societal expectations of privacy.

## *VII. Conclusions*

This Essay has described several essential elements for the development of a new ethical framework for big data research. A framework that is well-suited to the distinct and evolving features of big data research will achieve more appropriate privacy protection, enable greater harmonization of oversight across types of big data research, and facilitate the conduct of ethical research. Such a framework can catalyze big data utilization and help harness the tremendous value of big data in a sustainable and trust-building manner.